Multi-Modal Framework for Autism Severity Assessment Using Spatio-Temporal Graph Transformers

Kush Gupta[a], Amir Aly[b] and Emmanuel Ifeachor[c]
University of Plymouth, Plymouth, U.K.
{kush.gupta, amir.aly, E.Ifecahor}@plymouth.ac.uk

Keywords: Autism Spectrum Disorder, Severity, Spatio-Temporal Graph Transformer, Multi-Modal Data.

Abstract: Diagnosing Autism Spectrum Disorder (ASD) remains challenging, as it often relies on subjective evaluations and traditional methods using fMRI data. This paper proposes an innovative multi-modal framework that leverages spatiotemporal graph transformers to assess ASD severity using skeletal and optical flow data from the MMASD dataset. Our approach captures movement synchronization between children with ASD and therapists during play therapy interventions. The framework integrates a spatial encoder, a temporal transformer, and an I3D network for comprehensive motion analysis. Through this multi-modal approach, we aim to deliver reliable ASD severity scores, enhancing diagnostic accuracy and offering a scalable, robust alternative to traditional techniques.

1 INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects brain development, influencing how individuals perceive and engage with others, leading to challenges in social interaction and communication. It also involves repetitive and restricted patterns of behavior. The term ”spectrum” reflects the broad variety of symptom types and severity levels associated with ASD. The exact causes of ASD remain unclear, but research by (Lyall et al., 2017) indicates that both genetic and environmental factors are likely to contribute significantly. Traditional diagnostic techniques for ASD rely heavily on subjective behavioral evaluations (such as ADOS [1]), which can result in misidentification when distinguishing individuals with ASD from typically developing individuals (TC). Misdiagnoses are often linked to insufficient training and experience among medical professionals. These diagnostic challenges complicate pediatric screening efforts, as no straightforward diagnostic method is available. Accurate severity diagnoses of ASD often require continuous follow-up with patients to ensure dependable results.

Over the past two decades, structural MRI (sMRI) (Nickl-Jockschat et al., 2012) and resting-state functional MRI (rs-fMRI) (Santana et al., 2022) have been extensively utilized by researchers to develop machine learning models aimed solely at diagnosing autism rather than assessing its severity. Moreover, several challenges are associated with acquiring different types of MRI scans. First, MRI scanning is highly expensive, making large-scale data collection difficult. Second, individuals with autism spectrum disorder (ASD) often experience heightened anxiety, leading to discomfort while inside the MRI scanner. As a result, patients tend to move their heads during the scan, introducing noise into the data. Despite extensive pre-processing efforts, this noise remains difficult to eliminate, which affects the model's performance.

To resolve these challenges, one effective way is to use movement synchronization to assess ASD severity using intervention videos. Movement synchronization refers to the harmony of body gestures among interacting individuals, typically a therapist and an individual with ASD. Movement synchronization in psychological treatment could indicate a strong relationship between the individual and a psychologist (Nagaoka and Komori, 2008). Our technique examines movement synchronization between kids with ASD and therapeutic professionals during interac-

[a] https://orcid.org/0009-0008-9930-6435
[b] https://orcid.org/0000-0001-5169-0679
[c] https://orcid.org/0000-0001-8362-6292

[1] The Autism Diagnostic Observation Schedule (ADOS), developed by (Lord et al., 2000), is a partially structured diagnostic tool. It is employed to evaluate and determine the severity of autism spectrum disorder (ASD) through a standardized scoring system.

Gupta, K., Aly, A. and Ifeachor, E.
Multi-Modal Framework for Autism Severity Assessment Using Spatio-Temporal Graph Transformers.
DOI: 10.5220/0013236900003911

In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 2: HEALTHINF*, pages 686-693
ISBN: 978-989-758-731-3; ISSN: 2184-4305

tive therapy sessions. Based on the movement synchronization between the therapist and the child with ASD, an ASD severity score is assigned by our proposed framework.

Furthermore, an effective approach would involve using modalities such as skeletal data and optical flow, rather than directly processing the raw intervention videos. These modalities are preferred over raw RGB videos, as they provide more comprehensive information about body movements irrespective of the changes in the background and require less processing time. Optical flow refers to the perceived motion of individual pixels across two consecutive frames within an image plane. Extracted from raw video data, optical flow offers a compact representation of both the motion region and its velocity, enabling motion analysis without revealing personal identity . Skeleton data refers to a simplified representation of a human figure, capturing only key points (such as joints) of the body rather than full images or detailed appearances. These key points—such as the head, shoulder blades, elbows, wrists, hips, and knees are connected by lines that form a skeletal structure, representing the human body in a minimalistic way.

Hence, in our approach, we utilize the skeleton data, along with optical flow information, from the MMASD dataset (Li et al., 2023). This multimodal dataset is derived from interactive therapeutic interventions for children having autism. A total of 1,315 video clips were collected from 32 children with autism. Each sample includes three modalities taken from the raw videos: optical flow, 2D skeleton, and 3D skeleton. Also, the clinician evaluated the individuals with ASD and provided ADOS-2 scores for each child. Our proposed architecture is structured as an ensemble network comprising two primary branches to process these different modalities as demonstrated in Figure (1b). The first branch (shown in the dotted blue rectangle) consists of a spatial encoder and a temporal transformer to process skeletal data, while the second branch (dotted orange rectangle) utilizes a 3D convolutional network to incorporate optical flow information. Finally, a multilayer perceptron (MLP) serves as the classifier head, providing an autism severity score based on the child's performance when compared to the therapist during the intervention activity.

In summary, our approach will focus on utilizing modalities such as optical flow and skeletal data, in contrast to sMRI and fMRI. Collecting sMRI and fMRI data from patients with ASD is challenging, as discussed earlier in this section. Our study uses an ensemble framework, a spatial transformer to encapsulate the local relationships between the body connections, while a temporal transformer encodes global interconnections across multiple frames. Also, we utilized the Temporal Similarity Matrix (TSM), which represents sequential data in a graph format. TSMs excel in the analysis of human movement because they are resilient to perspective alterations and have strong generalization skills (Sun et al., 2015). In our study, similarity is computed between two skeletal sequences, comparing the interacting child and the therapist. The proposed method is designed to be identity-agnostic while retaining essential body movement features necessary for motion analysis and understanding.

The structure of this paper is organized as follows: a brief review of related work in this field is provided in the next section. Section (III) discusses the dataset used in the study, and Section (IV) outlines the proposed architecture. Section (V) presents the discussion and future work, and Section (VI) presents the conclusions drawn from the proposed approach.

2 RELATED WORK

Currently, three reliable and standardized instruments are commonly employed for autism diagnosis: the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2000), the Autism Diagnostic Interview-Revised (ADI-R) (Le Couteur et al., 1989), and the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) (American Psychiatric Association et al., 2013). While effective, these tools require considerable time for administration and score interpretation, causing delays in early intervention.

As a potential solution, machine learning techniques have been applied to develop classification models using rs-fMRI data. For example, a support vector classifier (SVC) was employed by (Abraham et al., 2017), achieving an accuracy of 67%. Similarly, (Monté-Rubio et al., 2018) utilized an SVC on the rs-fMRI dataset, attaining an accuracy of 62%. Over the past decade, advanced techniques such as deep neural networks (DNNs), long-short-term memory networks (LSTMs), and spatial-temporal graph transformers have gained traction in diagnosing ASD using rs-fMRI data. For example, (Sherkatghanad et al., 2020) developed a CNN-based classifier, achieving an accuracy of 70.22%. Furthermore, (Deng et al., 2022) proposed a linear spatial-temporal attention model to extract spatial and temporal representations to differentiate ASD subjects from typical controls using rs-fMRI data. However, these methods faced challenges in effectively extracting critical features from complex fMRI images, which limited their performance. Additionally, as discussed earlier, there are inherent difficulties associated with acquiring MRI scans.

Researchers have recently used movement synchrony methods to develop models for diagnosing Autism Spectrum Disorder. These methods are categorized into statistical approaches, which rely on low-level pixel features, and deep learning approaches, which extract high-level semantic information from video frames. Statistical methods (Altmann et al., 2021);(Tarr et al., 2018) have lately been widely used to facilitate the assessment of movement synchronization. It translates the raw video recordings of intervention into pixel-level presentations that capture temporal dynamics. The ultimate synchronization score is then calculated based on the correlation between the pixel sequences of different participants. However, these statistical methods are highly susceptible to noise, as they treat all pixels equally, which can lead to inaccuracies, especially in recordings from non-stationary cameras with dynamic backgrounds. Motion energy-based methods, among the most widely used statistical techniques (Altmann et al., 2021); (Tarr et al., 2018), require a predefined and fixed region of interest (ROI) and are limited in effectiveness if participants move outside this ROI. Moreover, these methods overlook the topological relationships among different human body parts. These limitations have led to poor performance and a lack of scalability for statistical methods when applied in specific contexts.

In contrast, deep learning methods have recently gained prominence for addressing the limitations of statistical approaches, showing enhanced performance in tasks related to human activity recognition (Dwibedi et al., 2020); (Zheng et al., 2021). Deep learning methods can leverage semantic information more effectively than statistical techniques, largely due to their capabilities of extracting characteristic features. (Calabrò et al., 2021) used a convolutional autoencoder to reconstruct inter-beat interval (IBI) segments from electrocardiogram (ECG) data. Their study focused on interactions between children with ASD and their counselors. A multi-task framework was introduced by (Li et al., 2021) to combine motion synchronization estimation with secondary tasks, such as interventional activity detection and action quality assessment (AQA). Nevertheless, both studies (Calabrò et al., 2021); (Li et al., 2021) required access to the raw video footage, limiting their ability to protect privacy.

However, our approach utilizes the MMASD dataset (Li et al., 2023), a privacy-focused dataset that derives skeletal (2D and 3D) and optical flow data from intervention videos. For processing these modalities, an ensemble network is proposed, employing the ST-GCN (Yan et al., 2018) to extract spatial and temporal features from the skeleton data of interacting individuals in the sequence of frames. Furthermore, it derives the features that represent the topological associations between the body joints, which helps the model to understand the body posture better. Additionally, an I3D (Carreira and Zisserman, 2017) model is used to analyze the optical flow information. The optical flow data in temporal and spatial dimensions enables the model to capture motion patterns of the interacting individuals in the frames over time. This additional information improves the model's ability to understand therapist and kid motions across frames appropriately. The proposed framework's ability to generalize is enhanced through the combination of multiple modalities. Scalability can be achieved by incorporating additional branches to process new modalities in the future. More details about the proposed approach are provided in section (4).

3 DATASET

The MMASD (Li et al., 2023) dataset includes a cohort of 32 children diagnosed with ASD, comprising 27 males and 5 females. Before participation, the Social Communication Questionnaire (SCQ) (Srinivasan et al., 2016) was utilized for initial screening, with final eligibility established through the ADOS scores and clinical evaluation. All recruited children were between the ages of 5 and 12 years. All videos were filmed in a domestic setting, where the video recorder focused on the area where each child participant engaged in their activities. This dataset comprises 1,315 video clips sourced from intervention recordings. There were three distinct themes: (1) Robot: children observed and imitated the movements demonstrated by a robot; (2) Rhythm: children and therapists engaged in therapeutic activities involving singing or playing musical instruments together; and (3) Yoga: children followed yoga exercises led by therapists, which included activities such as stretching, twisting, balancing, and similar movements. Based on the specific intervention activity, the data has been organized into eleven activity classes under these three primary themes, as outlined in the Table (1).

Three Key features were extracted from the original footage to retain essential movement details.

1. Optical flow: Optical flow refers to the perceived movement of objects within a scene, created by

Table 1: An overview of the 11 activity classes in the MMASD dataset (Li et al., 2023).

Theme	Activity Class	Count	Activity Description
Robotic	Arm swing	105	The participant lifts their left and right arms sequentially while maintaining a standing stance.
	Body swing	119	The body swings from left to right, with both hands outstretched, one hand behind the other.
	Chest expansion	114	The participant slowly expanded and contracted the chest.
	Squat	101	The participant assumes a crouching position with their knees flexed and maintains this posture in a repetitive manner.
Music	Drumming	168	The snare or Tubano drum is played by the participant using one or both hands.
	Maracas forward shaking	103	The participant actively shakes maracas, a percussion instrument frequently used in Caribbean and Latin music.
	Maracas shaking	130	The participant moves the maracas side to side in front of their chest.
	Sing and clap	113	Seated on the ground, the participant engages in singing and clapping simultaneously, an activity often performed at the beginning or conclusion of an intervention.
Yoga	Frog pose	113	The participant places their feet such that their big toes meet and opens their knees as far as possible.
	Tree pose	129	The participant assumes a tree pose, balancing on one leg while positioning the sole of the other foot against the inner thigh, calf, or ankle of the standing leg.
	Twist pose	120	Seated with legs crossed, the participant rotates their torso to one side while maintaining stability in the lower body.

the relative motion between the observer and the environment. They used the Lucas-Kanade method (Lucas and Kanade, 1981) to obtain optical flow information, a widely used technique in computer vision for estimating object movement between frames. This method assumes minimal, consistent displacement in image content within a neighborhood around a given point.

2. 2D skeleton: Skeletal data has significant benefits over raw RGB data since it simply contains the locations of human body joints on a 2D plane, and provides context-agnostic information. This data enables models to focus on robust features of bodily movements.

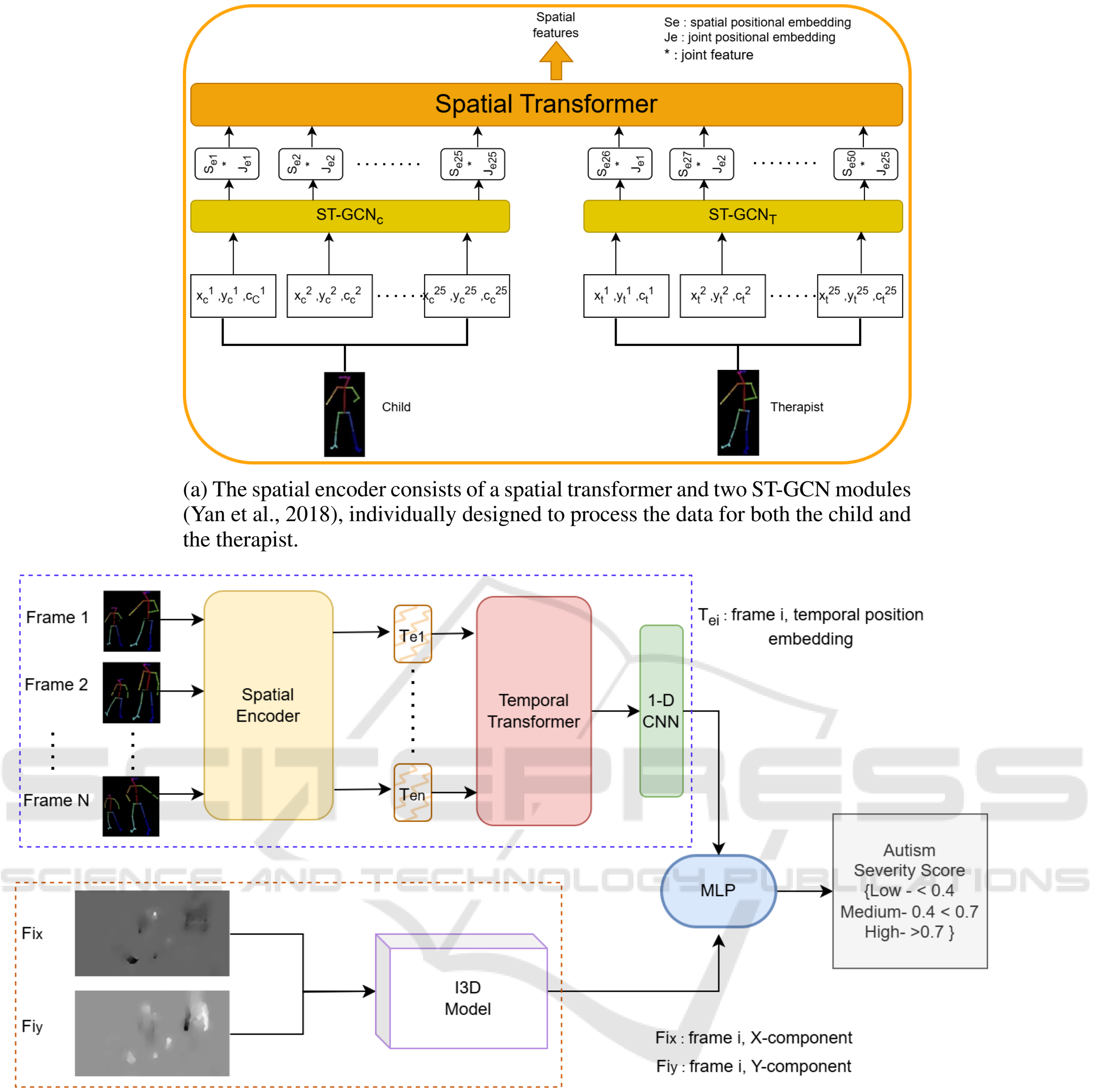
   In the MMASD dataset, 2D skeletons were extracted from recorded videos using OpenPose (Cao et al., 2017). This library detects key human structural points, including joints and body components, in real-time from images or video frames for multiple users simultaneously. Confidence scores for each body component were generated, followed by association with individual persons using Part Affinity Fields.

3. 3D skeleton: 3D skeletons represent each key joint in three-dimensional coordinates, adding a depth dimension. For extracting 3D skeleton data, the Regression of Multiple 3D People (ROMP) method, introduced by (Sun et al., 2021), was applied, providing depth and pose estimation from single 2D images. The approach estimates various differentiable maps from the picture, including a heatmap for the body center and a map for the mesh parameters. These maps are used to produce 3D body mesh parameter vectors for each person via parameter sampling. These vectors are processed through the Skinned Multi-Person Linear Model (SMPL) model to generate multi-person 3D models.

Demographic information and autism assessment results for all the children were also reported, encompassing details such as motor functioning scores, date of birth, and autism spectrum disorder severity levels using the ADOS-2 scores. Even though, the MMASD dataset contains various modalities, including 2D and 3D skeletal data as well as optical flow information, the relatively limited number of data samples may restrict the model's adaptability in different real-world scenarios.

4 PROPOSED ARCHITECTURE

Figure (1b) illustrates the structure of the suggested framework. It comprises four main elements. First is the Spatial Encoder, which includes the ST-GCN and a spatial transformer to derive spatial information between the interacting individual with ASD and the therapist from each frame refer to Figure (1a). Second, is the Temporal Transformer designed to extract temporal features for the entire sequence. Third is the I3D model, a flow-stream Inflated 3D ConvNet (I3D) was employed to process the optical flow data, and finally, a multi-layer perception (MLP) is the classification head responsible for predicting the autism severity score (Hus et al., 2014). This section provides a brief overview of these main components and a detailed explanation of the proposed methodology.



(a) The spatial encoder consists of a spatial transformer and two ST-GCN modules (Yan et al., 2018), individually designed to process the data for both the child and the therapist.

(b) The top branch (dashed blue rectangle) processes the skeleton data, and the bottom (dashed orange rectangle) processes optical flow.

Figure 1: Illustration of the basic building blocks of the proposed model for autism spectrum disorder (ASD) prediction.

4.1 Spatial Encoder

The spatial encoder as shown in the Figure (1a), is composed of two primary components: Spatio-Temporal Graph Convolutional Networks (ST-GCN) and a spatial transformer. Specifically, two separate ST-GCNs are employed—one dedicated to processing the child's skeletal data and the other for handling the therapist's data. The skeleton data was generated by a pose detector, where every joint J_i is represented as a 3-D vector (x_i, y_i, c_i). Here, x_i and $y_i \in \mathbb{R}^2$, as they denote the coordinates of joint J_i, and $c_i, \in [0, 1]$, represents the joint confidence values estimated by the pose detector. For each joint J_i, a hidden embedding is produced through the $ST - GCN_k$ where k $\in$ to [child, therapist], which then serves as the input to the spatial transformer.

4.1.1 ST-GCN

ST-GCN models the spatial and temporal structure of skeleton data by using a graph convolutional network designed for skeleton-based data analysis. Each skeleton joint is represented as a node, while connections between adjacent joints are represented as edges. Given a set of vertices V, where each vertex

$v_i \in V$ represents a specific joint, and edges based on natural human joint connections, ST-GCN applies graph convolutions to capture the spatial dependencies between these joints. Mathematically, the feature map of each vertex v_i before the convolution can be represented as $f_in(v_i)$. After applying the ST-GCN convolution, the output feature map $f_out(v_i)$ is obtained using:

$$f_{out}(v_i) = \sum_{B \in B(i)} \sum_{v_i \in B(i)} \frac{1}{|B|} f_{in}(v_j).w(B(i)) \quad (1)$$

here $B(i)$ denotes the neighborhood of v_i, determined by both human body connections and a predefined partition rule. The function $w(B(i))$ represents the weights assigned to each neighborhood, and $|B|$ is the normalization factor for the neighborhood size.

4.1.2 Spatial Transformer

To adjust the unique structure of dyadic (child-therapist) interactions by introducing a spatial transformer that generates spatial embeddings, we incorporated both the joint's position and its correspondence across interacting individuals. The spatial transformer combines three main components to form a spatial feature vector: (1) patch embedding, (2) spatial positional embedding, and (3) joint index embedding shared between matching joints in child-therapist pairs.

The final spatial embedding S is computed by:

$$S = ST(P + E_{pos} + E_{joint}) \quad (2)$$

where P represents the patch embedding, E_{pos} is the spatial positional embedding that maintains the joint order, and E_{joint} is the joint index embedding, enabling the transformer to attend to the corresponding joints across the child and the therapist in synchrony assessments.

4.2 Temporal Transformer

The Temporal Transformer captures the temporal relationships across frames, enhancing the model's ability to assess movement synchrony. It applies attention to the sequential frames and includes a temporal similarity matrix (TSM), which captures periodic movements inherent in autism intervention sessions. The temporal embedding includes: (1) Patch embedding P', which is the output from the spatial transformer (2) temporal positional embedding E'_{pos}, which preserves the temporal order of the frames, (3) and frame uncertainty embedding $E_{uncertainity}$, representing the confidence score for the frame, calculated from joint confidence scores. The frame uncertainty embedding $E_{uncertainity}$ for a specific frame t is computed by:

$$E_{uncertainity} = Linear(c_1^1, c_1^2.., c_1^{25}; c_2^1, c_2^2.., c_2^{25}) \quad (3)$$

where each c^i represents the confidence score of a joint. To account for the periodic nature of therapeutic interventions, a temporal self-similarity matrix S (Dwibedi et al., 2020) is integrated into the computation of temporal attention. S is represented as a square matrix M^2 , where M denotes the number of frames is sequence. Each element $M[i][j]$ represents the resemblance between the pose X_1^i from the first individual at timestamp i and the pose X_2^j from the second individual at timestamp j.

Rather than calculating similarity directly from coordinates $((x_i, y_i))$, the computation is based on corresponding d-dimensional feature vectors f_{d1}^i and f_{d2}^j, produced by the ST-GCN. The similarity function is formulated by calculating the Euclidean distance between these feature vectors. subsequently, a softmax operation is applied along the time axis. The temporal similarity matrix M defined between corresponding feature vectors of child and therapist frames:

$$M[i,j] = \frac{-1}{d} \sqrt{\sum_{m=1}^{d} \left\| f_1^i[m] - f_2^j[m] \right\|^2} \quad (4)$$

here, M is processed through a convolutional layer to produce a feature map $\hat{M}$, which is subsequently included in the computation of temporal attention.

4.3 I3D

A common approach to understanding human activities in videos is to utilize 3D convolutional neural networks, which apply the convolutional operation over the spatiotemporal sequence of frames. These frames represent the motion information between consecutive video frames. In our approach to processing optical flow data, our model is based on the optical flow stream of the Inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017). Additionally, we will employ the model previously trained on the Kinetics dataset (Kay et al., 2017). The model employs 3D convolutions to process the optical flow data in temporal and spatial dimensions, enabling it to capture motion patterns over time. Solving the optical flow equation across a window centered on the point effectively captures the interacting individuals' motion in image sequences. The optical flow stream delivers explicit motion information of the interacting individuals, assisting the model in comprehending the movement of the therapist and child across frames.

4.4 MLP

The classification head consists of a Multi-Layer Perceptron (MLP). Given that the autism spectrum has been structured as a three-class (low, medium, high) classification problem. The proposed model will be trained using cross-entropy loss, calculated between the model prediction score and the corresponding ADOS-2 score of the sample. The final result would be a probability score that corresponds to the severity level of autism.

5 DISCUSSION AND FUTURE WORK

To validate the model, the provided ADOS-2 scores will be utilized. The final probability score provided by the model will be compared to these reference scores to determine its performance. Future work could enhance the model by incorporating additional modalities like 3D Body mesh, further improving diagnostic accuracy and expanding its usability in diverse settings. Moreover, a dataset with a larger number of samples and different activity classes beyond those currently available in the MMASD dataset would greatly enhance the model's robustness and adaptability across a range of real-world scenarios.

6 CONCLUSIONS

Gathering sMRI and fMRI data poses major hurdles, including high operating expenses and the discomfort experienced by individuals with ASD within the scanner. This discomfort often introduces intrinsic noise into the data, making it difficult to fully eliminate, even with extensive pre-processing efforts. Hence, in our study we propose a multi-modal framework that combines other modalities like skeletal and optical flow data for ASD diagnosis, to analyze the movement synchronization between children and therapists. By using the spatio-temporal Graph Convolution Neural Network (ST-GCN) and spatial-temporal graph transformers, this model effectively captures spatial and temporal dynamics essential for ASD intervention assessment. More specifically, the Spatial-Temporal Graph Convolutional Neural Network (ST-GCN) leverages the body's intrinsic connections to better depict joint topology. Further, the model's integration of a temporal similarity matrix improves its robustness in various therapeutic activities. Additionally, optical flow data is utilized to effectively capture motion patterns between the child with ASD and the therapist over time. This supplementary information enables the model to more accurately interpret the movements of both the therapist and child across frames. Ultimately, the model outputs an autism severity score, providing valuable insights for therapists to take further action. Although the MMASD dataset includes different modalities, such as 2D and 3D skeletal data and optical flow information, the number of data samples is relatively small, which may limit the model's adaptability in various real-world circumstances.

REFERENCES

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., and Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, 147:736–745.

Altmann, U., Brümmel, M., Meier, J., and Strauss, B. (2021). Movement synchrony and facial synchrony as diagnostic features of depression: A pilot study. *The Journal of nervous and mental disease*, 209(2):128–136.

American Psychiatric Association, D., American Psychiatric Association, D., et al. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Calabrò, G., Bizzego, A., Cainelli, S., Furlanello, C., and Venuti, P. (2021). M-ms: A multi-modal synchrony dataset to explore dyadic interaction in ASD. *Progresses in Artificial Intelligence and Neural Systems*, pages 543–553.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Real-time multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Deng, X., Zhang, J., Liu, R., and Liu, K. (2022). Classifying ASD based on time-series fMRI using spatial-temporal transformer. *Comput. Biol. Med.*, 151(Pt B):106320.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. (2020). Counting out time: Class agnostic video repetition counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10387–10396.

Hus, V., Gotham, K., and Lord, C. (2014). Standardizing ADOS domain scores: separating severity of social affect and restricted and repetitive behaviors. *J. Autism Dev. Disord.*, 44(10):2400–2412.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Le Couteur, A., Rutter, M., Lord, C., Rios, P., Robertson, S., Holdgrafer, M., and McLennan, J. (1989). Autism diagnostic interview: a standardized investigator-based instrument. *J. Autism Dev. Disord.*, 19(3):363–387.

Li, J., Bhat, A., and Barmaki, R. (2021). Improving the movement synchrony estimation with action quality assessment in children play therapy. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 397–406.

Li, J., Chheang, V., Kullu, P., Brignac, E., Guo, Z., Bhat, A., Barner, K. E., and Barmaki, R. L. (2023). MMASD: A Multimodal Dataset for Autism Intervention Analysis. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 397–405, New York, NY, USA. Association for Computing Machinery.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30:205–223.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679.

Lyall, K., Croen, L., Daniels, J., Fallin, M. D., Ladd-Acosta, C., Lee, B. K., Park, B. Y., Snyder, N. W., Schendel, D., Volk, H., et al. (2017). The changing epidemiology of autism spectrum disorders. *Annual review of public health*, 38(1):81–102.

Monté-Rubio, G. C., Falcón, C., Pomarol-Clotet, E., and Ashburner, J. (2018). A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods. *Neuroimage*, 178:753–768.

Nagaoka, C. and Komori, M. (2008). Body movement synchrony in psychotherapeutic counseling: A study using the video-based quantification method. *IEICE transactions on information and systems*, 91(6):1634–1640.

Nickl-Jockschat, T., Habel, U., Maria Michel, T., Manning, J., Laird, A. R., Fox, P. T., Schneider, F., and Eickhoff, S. B. (2012). Brain structure anomalies in autism spectrum disorder—a meta-analysis of VBM studies using anatomic likelihood estimation. *Human Brain Mapping*, 33(6):1470–1489.

Santana, C. P., de Carvalho, E. A., Rodrigues, I. D., Bastos, G. S., de Souza, A. D., and de Brito, L. L. (2022). rs-fMRI and machine learning for ASD diagnosis: a systematic review and meta-analysis. *Scientific Reports*, 12(1):6030.

Sherkatghanad, Z., Akhondzadeh, M., Salari, S., Zomorodi-Moghadam, M., Abdar, M., Acharya, U. R., Khosrowabadi, R., and Salari, V. (2020). Automated detection of autism spectrum disorder using a convolutional neural network. *Frontiers in neuroscience*, 13:1325.

Srinivasan, S. M., Eigsti, I.-M., Neelly, L., and Bhat, A. N. (2016). The effects of embodied rhythm and robotic interventions on the spontaneous and responsive social attention patterns of children with autism spectrum disorder (ASD): A pilot randomized controlled trial. *Research in autism spectrum disorders*, 27:54–72.

Sun, C., Junejo, I. N., Tappen, M., and Foroosh, H. (2015). Exploring sparseness and self-similarity for action recognition. *IEEE Transactions on Image Processing*, 24(8):2488–2501.

Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M. J., and Mei, T. (2021). Monocular, one-stage, regression of multiple 3D people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188.

Tarr, B., Slater, M., and Cohen, E. (2018). Synchrony and social connection in immersive virtual reality. *Scientific reports*, 8(1):3693.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., and Ding, Z. (2021). 3D human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11656–11665.